

IMAGE RETRIEVAL WITH HIERARCHICAL MATCHING PURSUIT

Shasha Bu, Yu-Jin Zhang

Department of Electronic Engineering,
Tsinghua University,
Beijing 100084, China

Email: boss12@mails.tsinghua.edu.cn, zhang-yj@mail.tsinghua.edu.cn

ABSTRACT

A novel representation of images for image retrieval is introduced in this paper, by using a new type of feature with remarkable discriminative power. Despite the multi-scale nature of objects, most existing models perform feature extraction on a fixed scale, which will inevitably degrade the performance of the whole system. Motivated by this, we introduce a hierarchical sparse coding architecture for image retrieval to explore multi-scale cues. Sparse codes extracted on lower layers are transmitted to higher layers recursively. With this mechanism, cues from different scales are fused. Experiments on the Holidays dataset show that the proposed method achieves an excellent retrieval performance with a small code length.

Index Terms— CBIR, sparse coding, hierarchical matching pursuit, bag-of-features

1. INTRODUCTION

Image retrieval has been increasingly popular in recent years. Searching images such as pictures of a scenic spot or an animal has become a part of everyday life for many people, either from the internet or database in hand. However, with image database growing increasingly larger, how to find the intended images from so many images is a problem presented in image retrieval. A lot of works have been done in this field [1][2][3][4][5].

Recent works on image retrieval mainly concentrate on content based image retrieval (CBIR). Features from images are extracted and compared for similarity measurement based on which the most similar images to the query are returned.

Bag-of-features (BoF) model [6] is extensively used in CBIR which often obtains good performance. Methods following such a framework often use Scale-invariant feature transform (SIFT) [7], which is robust against many image transformations. However, the vector quantization (VQ) [8]

in BoF model only assumes that each feature is related to a single visual word, and thus ignores the correlation between the feature and other words. What is more, SIFT is a local feature which is unable to capture the global cues. And features of the same image are irrelevant to each other, limiting the fusion of cues between them. Sparse coding techniques and global features have been proposed to fix the problem [9][10][11][12][13][14][15]. Nevertheless, neither utilizing one-layer sparse coding nor leveraging global feature on a fixed scope can cues of different scales be adequately explored. The success of hierarchical matching pursuit (HMP) algorithm in classification [16] motivates us to employ the hierarchical sparse coding architecture in image retrieval to explore multi-scale cues.

A global feature using HMP is introduced in this paper for image retrieval, which has not been considered in this field to our knowledge. The global cues as well as features on different scales are extracted, forming a sparse representation. Images are first partitioned into patches of different sizes. Then, sparse codes are extracted from smaller patches and spatially pooled on larger patches recursively. Finally, a hierarchical sparse coding architecture is constructed, and sparse representations extracted from the hierarchical layers are adopted for retrieval. Experiments conducted on the Holidays dataset [17] demonstrate the effectiveness of the proposed approach, where excellent performance compared with prior methods is obtained.

2. SPARSE CODING IN CBIR

This section presents the procedure of utilizing sparse coding for CBIR. A standard sparse coding model can be formulated as follows. Given an over completed codebook \mathbf{C} ($\mathbf{C} \in \mathbb{R}^{D \times K}$) and a basic feature \mathbf{y} ($\mathbf{y} \in \mathbb{R}^D$), a vector \mathbf{x} ($\mathbf{x} \in \mathbb{R}^K$) with sparsity L is generated to approximate \mathbf{y} [11] as

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{C}\mathbf{x}\|^2, s.t. \|\mathbf{x}\|_0 \leq L. \quad (1)$$

Orthogonal matching pursuit (OMP) [16] is usually employed to solve Eq. (1).

This work was supported by National Nature Science Foundation (NNSF: 61171118) and Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP-20110002110057).

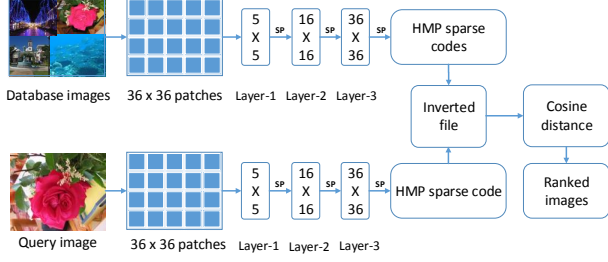


Fig. 1. Architecture of a three-layer hierarchical matching pursuit. Spatial max pooling is denoted by SP.

When sparse coding is used in CBIR, features are extracted from the image and sparsely coded using Eq. (1). Then, max-pooling [16] is applied to all sparse codes of the image to form a sparse representation which is used for similarity measurement in the search step.

The BoF model can also be treated as a special case of sparse representation [18]. Low-level features extracted from the image are quantized to the nearest visual words in the codebook using VQ as

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{C}\mathbf{x}\|^2, \text{ s.t. } \|\mathbf{x}\|_0 = 1, \|\mathbf{x}\|_1 = 1, \mathbf{x}(i) \geq 0, \forall i. \quad (2)$$

Codes of all features of an image are aggregated using average pooling [9], generating a final sparse representation of the BoF model. Note that Eq. (2) only allows a sparsity level 1 of vector \mathbf{x} which means a feature is assigned to only one visual word in the codebook in a hard manner. However, this may not be appropriate since a feature could also be related to multiple visual words, which has been proved in [9], and thus the retrieval performance of BoF is limited while OMP can be utilized to improve it by assigning a feature to more visual words.

3. PROPOSED APPROACH

This section describes the hierarchical matching pursuit for image retrieval approach (HMP-IR). The correlations with multiple visual words are explored using OMP, and discriminative features of different scales are extracted using hierarchical sparse coding layers. Global cues can also be utilized by max pooling on spatial pyramids. A three-layer architecture of the whole HMP-IR algorithm is shown in Fig. 1. We use the same parameter settings as [16]. More details are shown below.

3.1. Extracting HMP Representation

This subsection shows how to form a sparse HMP representation for a given image. The HMP representation consists of multiple layers. Input data of the first layer are raw patches

sampled from images, and input of the higher are the pooled sparse codes from the previous layer. Sparse codes are extracted and pooled recursively on different layers. Mutual incoherence KSVD (MI-KSVD) method is adopted for codebook training [16]. A spatial pyramid is constructed on the final layer. The coding procedure for a three-layer HMP-IR is as follows.

The first layer: Sparse codes from small patches are extracted and adopted for generating representations for mid-level patches. A mid-level patch P (e.g. 16x16) is further divided into small spatial cells, and each cell is divided into small image patches (e.g. 5x5) with overlaps. A sparse code is extracted from each small patch using the codebook of this layer. Codes of small patches within a cell Ce are aggregated using max-pooling as

$$F(Ce) = \max_{j \in Ce} [\max(x_{j1}, 0), \dots, \max(x_{jM}, 0), \max(-x_{j1,0}), \dots, \max(-x_{jM,0})], \quad (3)$$

where j is the index of a small patch within the cell Ce , and x_{jM} is the M -th element of the j -th sparse code vector x_j in cell Ce . The positive and negative elements of vector x_j are split into separate features and weighted differently by the higher layer encoder. Feature F_P of mid-level patch P is the concatenation of codes of all spatial cells Ce_s^P , $s = [1, 2, \dots, S]$ in P as

$$F_P = [F(Ce_1^P), \dots, F(Ce_2^P), \dots, F(Ce_S^P)]. \quad (4)$$

The feature F_P is then ℓ_2 -normalized [16] and fed to the second layer.

The second layer: The features F_P from the first layer are delivered to the second layer and processed the same way as raw patches on the first layer. Sparse codes for each feature F_P are drawn and spatially max-pooled within each cell. Codes of each cell are concatenated on large image patches (e.g., 36x36). Then, the concatenated features on large image patches are normalized and transmitted to the third layer.

The third layer: The features generated from the second layer are sparsely coded on the third layer. On this final layer, max pooling on spatial pyramids on the whole image is conducted. The pooled descriptors are ℓ_2 normalized to form a sparse representation for the whole image. The coding procedures for the three different HMP-IR methods are illustrated in Fig. 2.

3.2. HMP Representation for Image Retrieval

Representations of the database images computed in Sec. 3.1 are sparse, and are utilized for generating an inverted file [6] to speed up the searching procedure. In the search step, the query is coded in the same way, and then the inverted file is used to identify the candidate images. Cosine distance [6] is employed to evaluate the similarities between the candidates and the query.

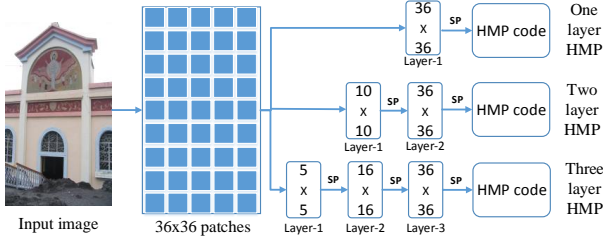


Fig. 2. Procedures of three different HMP-IR methods. SP indicates spatial max pooling.

4. EXPERIMENTS

In this section, performances of the proposed approach utilizing different numbers of layers are presented. Comparisons with the BoF model utilizing RootSIFT features [19] and other image retrieval methods are conducted on different code lengths. RootSIFT features are produced from SIFT features and perform better than the latter. The mean Average Precision (mAP) is adopted to evaluate different methods.

4.1. Parameter Settings

Two groups of HMP-IR methods are utilized to evaluate the performance on the Holidays dataset [17] with three different numbers of layers. In each group, one-layer HMP-IR (HMP-IR1), two-layer HMP-IR (HMP-IR2) and three-layer HMP-IR (HMP-IR3) methods are implemented on 36x36 image patches. Codebook sizes of each group on the final layer are set to 500 and 1000, respectively, to test the influence of codebook size on retrieval performance.

On the final layer, image-level features are obtained by max pooling on spatial pyramids on the whole image. Parameters of spatial pyramids are set to 1x1, 2x2 and 3x3 on the whole image. Different combinations of them are implemented. Note that the length of descriptor before spatial max pooling is double the size of the codebook on the final layer because of pooling in Eq. (3).

We adopt the BoF model [20] as baseline. An ℓ_p -norm inverse document frequency (IDF) [20] weighting strategy ($p = 3$) is employed to obtain a higher result.

4.2. Retrieval Results on the Holidays Dataset

The Holidays dataset is widely used in image retrieval and contains 1491 color images taken on a large variety of scenes with 500 queries [17]. A few example images are shown in Fig. 3.

Comparison of the proposed HMP-IR2 method (pooled on 1x1 pyramid) with BoF and other state of art methods such as vector of locally aggregated descriptors (VLAD) [10] and Fisher [12] is presented in Table 1. Codebook size is denoted



Fig. 3. A few examples on the Holiday dataset.

Table 1. Comparison of different methods on the Holidays dataset.

Methods	K	D	mAP
BoF[20]	20 000	20 000	0.4713
VLAD[10]	64	8192	0.526
Fisher[12]	64	4096	0.595
HMP-IR2	1000	2000	0.6822

Table 2. Performances of three HMP-IR methods with different codebook sizes (K) on 1x1 pyramid.

mAP	HMP-IR1	HMP-IR2	HMP-IR3
$K = 500$	0.4849	0.6537	0.6390
$K = 1000$	0.4992	0.6882	0.6603

by K . The final length of the feature is denoted by D . Results from Table 1 show that the HMP-IR method outperforms the others with a shorter code. The storage is reduced from 365MB to 6.63MB compared with BoF. Query time for each method are 0.0587s and 0.0554s, respectively. The query time doesn't decrease because a single feature is assigned to more visual words in HMP-IR, and thus more candidates are selected for similarity measurement.

As is shown in Fig. 4, the HMP-IR2 extracts discriminative features from multiple scales (the small-scale blue river and the grass land and mountain of large scale), while BoF mainly learns features of fixed scale (the large-scale white road and mountain) which take more area of the image than others.

Performances of the two groups of HMP-IR methods are shown in Table2 with max pooling on 1x1 pyramid. The final codebook sizes (K) are 500 and 1000, respectively.

According to Table 2, performance is improved with a larger codebook since more cues can be encoded. HMP-IR2 and HMP-IR3 outperform HMP-IR1 which proves that the correlations between visual words are excavated by delivering codes between different hierarchical layers, and cues of image are thoroughly used, which is shown in Fig. 4. Performance of three-layer HMP-IR is not as good as two-layer



Fig. 4. The top 8 images returned by HMP-IR2 and BoF. The first and second rows correspond to HMP-IR2, and the lower rows to BoF. Incorrect results are marked with red boxes. The mAP for each are 0.8012 and 0.0616, respectively. Number of ground truth is 8.



Fig. 5. Three failure examples of HMP-IR3 ($K = 1000$), on 1×1 pyramid. Queries and the corresponding ground truths are shown in each group.

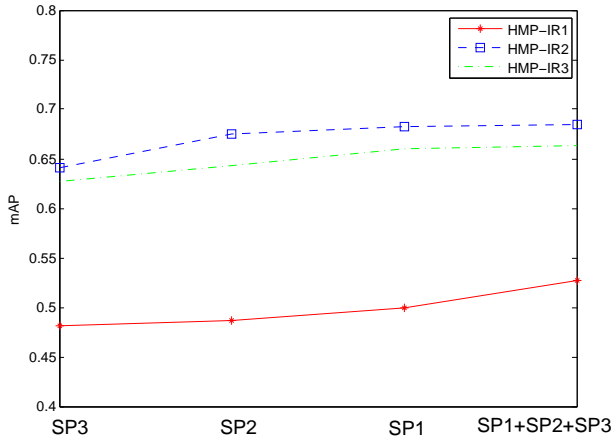


Fig. 6. Performance of three HMP-IR methods ($K = 1000$) on different pyramids. SP1, SP2, SP3 indicates a pyramid scale of 1×1 , 2×2 and 3×3 , respectively. SP1+SP2+SP3 denotes the combination of three pyramids.

HMP-IR. A few failure cases of HMP-IR3 with 1000 code-book size are shown in Fig. 5, where the ground truths are of different view points from the queries. This may indicate that angle cues are lost through too many layers of sparse coding.

Fig. 6 shows the performance of the second group of HMP-IR methods on different pyramids. SP1, SP2 and SP3

denotes 1×1 , 2×2 and 3×3 spatial pyramid on the whole image, respectively. SP1+SP2+SP3 indicates the combination of three pyramids. It can be drawn from Fig. 6 that better performance is obtained on larger grid (e.g., 1×1) which is easy to understand as pooling on larger grid can embed more spatial cues.

5. CONCLUSION

In this paper, we introduce the hierarchical matching pursuit method from image classification and modify the procedure to apply it to image retrieval. Multi-scale features are fused, and global cues are explored to obtain a better performance. Experiments show that our approach outperforms many other methods with a shorter descriptor. Future works include testing the scalability on large scale and different datasets and fusion with other features.

6. REFERENCES

- [1] L. Zheng and S. Wang, “Visual phraselet: Refining spatial constraints for large scale image search,” *IEEE Signal Process. Lett.*, vol. 20, no. 4, pp. 391–394, 2013.
- [2] L. Zheng, S. Wang, P. Guo, H. Liang, and Q. Tian, “Bayes merging of multiple vocabularies for scalable image retrieval,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.
- [3] Z. Liu, S. Wang, L. Zheng, and Q. Tian, “Visual reranking with improved image graph,” in *ICASSP*, 2014, pp. 6909–6913.
- [4] L. Zheng, S. Wang, Z. Liu, and Q. Tian, “Packing and padding: Coupled multi-index for accurate image retrieval,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.
- [5] L. Zheng, S. Wang, F. He, and Q. Tian, “Seeing the big picture: Deep embedding with contextual evidences,” *arXiv preprint arXiv:1406.0132*, 2014.
- [6] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *IEEE International Conference on Computer Vision*. IEEE, 2003, pp. 1470–1477.
- [7] D. G. Lowe, “Object recognition from local scale-invariant features,” in *IEEE International Conference on Computer vision*. IEEE, 1999, vol. 2, pp. 1150–1157.
- [8] R. Gray, “Vector quantization,” *IEEE ASSP Magazine*, vol. 1, no. 2, pp. 4–29, 1984.
- [9] J. Shi, Z. Jiang, H. Feng, and L. Zhang, “Sift-based elastic sparse coding for image retrieval,” in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2012, pp. 2437–2440.
- [10] H. Jégou, M. Douze, C. Schmid, and P. Pérez, “Aggregating local descriptors into a compact image representation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3304–3311.

- [11] J. J. Thiagarajan, R. K. Natesan, P. Sattigeri, and A. Spanias, "Supervised local sparse coding of sub-image features for image retrieval," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2012, pp. 3117–3120.
- [12] F. Perronnin, Y. Liu, J. Sánchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3384–3391.
- [13] Y. Zheng, Y. Zhang, and H. Larochelle, "Topic modeling of multimodal data: an autoregressive approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.
- [14] B. Liu, Y. Wang, Y. Zhang, and Y. Zheng, "Discriminant sparse coding for image classification," in *ICASSP*. IEEE, 2012, pp. 2193–2196.
- [15] B. Liu, Y. Wang, Y. Zhang, and B. Shen, "Learning dictionary on manifolds for image classification," *Pattern Recognition*, vol. 46, no. 7, pp. 1879–1890, 2013.
- [16] L. Bo, X. Ren, and D. Fox, "Multipath sparse coding using hierarchical matching pursuit," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 660–667.
- [17] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *Computer Vision—ECCV 2008*, pp. 304–317. Springer, 2008.
- [18] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3360–3367.
- [19] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 2911–2918.
- [20] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Lp-norm idf for large scale image search," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 1626–1633.